Evaluating a shrinkage estimator for the treatment effect in clinical trials

Zwet, E.W. van; Tian, L.; Tibshirani, R.

Received: 2 April 2023 | Revised: 21 November 2023 | Accepted: 3 December 2023

DOI: 10.1002/sim.9992

RESEARCH ARTICLE

Statistics in Medicine WILEY

Evaluating a shrinkage estimator for the treatment effect in clinical trials

Erik W. van Zwet[1] | **Lu Tian**[2] | **Robert Tibshirani**[2,3]

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[2]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

[3]Department of Statistics, Stanford University, Stanford, California, USA

Correspondence
Erik W. van Zwet, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.
Email: E.W.van_Zwet@lumc.nl

The main objective of most clinical trials is to estimate the effect of some treatment compared to a control condition. We define the signal-to-noise ratio (SNR) as the ratio of the true treatment effect to the SE of its estimate. In a previous publication in this journal, we estimated the distribution of the SNR among the clinical trials in the Cochrane Database of Systematic Reviews (CDSR). We found that the SNR is often low, which implies that the power against the true effect is also low in many trials. Here we use the fact that the CDSR is a collection of meta-analyses to quantitatively assess the consequences. Among trials that have reached statistical significance we find considerable overoptimism of the usual unbiased estimator and under-coverage of the associated confidence interval. Previously, we have proposed a novel shrinkage estimator to address this "winner's curse." We compare the performance of our shrinkage estimator to the usual unbiased estimator in terms of the root mean squared error, the coverage and the bias of the magnitude. We find superior performance of the shrinkage estimator both conditionally and unconditionally on statistical significance.

KEYWORDS
clinical trial, Cochrane Review, shrinkage

1 | INTRODUCTION

The Cochrane collaboration is a global independent network that aims to gather and summarize the best evidence—usually randomized controlled trials or RCTs—from medical research. The Cochrane Database of Systematic Reviews (CDSR) contains the results of tens of thousands of clinical trials covering any topic relevant to health care, including health services. While there is evidence that the database may suffer from some publication bias and dubious research practices such as *p*-hacking,[1] it is arguably the largest, most comprehensive and most reliable collection of evidence in medicine currently available. See Reference 2 for a detailed description of the CDSR.

From a meta-scientific point of view, the CDSR is a unique resource to study how medical research is conducted. For example, in References 3-5 we studied the distribution of the power of the two-sided test for detecting the *true* effect in studies from CDSR. We found that the median power is only about 14% across the entire CDSR. Low statistical power against the true effect has been observed before in various domains of biomedical research,[6,7] and it is actually not that surprising. First of all, trials are designed to have good power against the minimal effect that is of clinical interest, *not* against the true effect (which is of course unknown). Other factors that may also contribute to low power are: limited financial resources, lack of time, difficulties with subject recruitment, or larger between-subject variation than anticipated.

The fact that the power against the true effect is often low has serious consequences for our inferences. One such consequence is the well-known "winner's curse," which is the tendency of statistically significant effects to be overestimated

while the associated confidence interval does not attain the nominal coverage. When the power is low, the winner's curse is especially severe.[8-10] A related problem is that replication studies of the same size will often fail to reach statistical significance.[5] In this article, we present a quantitative assessment of the performance of the usual, unbiased estimator on average across the trials in the CDSR in terms of the root mean squared error (RMSE), the coverage and the bias of the magnitude, both conditionally and unconditionally on statistical significance. We also consider the performance of a novel shrinkage estimator which we recently proposed in References 3 and 4.

The article is organized as follows. We describe our shrinkage estimator in Section 2; in the Appendix, we also provide a few lines of R code to compute it together with its SE. In our previous work, we used the CDSR merely as a large collection of individual trials when it is actually a collection of systematic reviews. In Section 3, we use this structure to make a synthetic "copy" of the CDSR. We then use this copy to evaluate and compare the performance of the usual unbiased estimator and our shrinkage estimator. We find superior performance of the shrinkage estimator on average over the trials of the synthetic CDSR both conditionally and unconditionally on statistical significance. We claim that the performance across the synthetic CDSR gives a good indication of the performance across the real CDSR. In Section 3.3, we provide additional direct support for the validity of this claim by means of cross-validation without simulating synthetic data.

We briefly elaborate on this cross-validation procedure. We naturally think of the unbiased and shrinkage estimators as estimators of the effect in a particular trial. However, we can also view them as estimators of the pooled effect in any meta-analysis that includes that trial. We can obtain a third estimator of the pooled effect by leaving out that one trial, and repeating the meta-analysis using only the remaining trials. This third estimator is unbiased and independent of the other two. We can now compare the unbiased and shrinkage estimators from the trial that was left out to the pooled estimator from the remaining trials.

This cross-validation approach is reminiscent of the well-known study by Efron and Morris to predict baseball batting averages on the basis of the first 45 at-bats.[11] The individual trials from the CDSR play the role of the batting averages over the first 45 at-bats, while the remaining studies in the same meta-analysis play the role of the batting averages over the remainder of the season.

In Section 4, we compare the estimators more finely, stratifying the trials by medical field. We end the article with a brief discussion in Section 5.

2 | DEFINING THE SHRINKAGE ESTIMATOR

We now discuss the shrinkage estimator which we proposed in References 3 and 4. For more detail, we refer to those papers. We represent an individual trial by a set of three numbers (β, b, s), where β is the primary efficacy parameter and b is an unbiased, normally distributed estimator with SE s, that is,

$$b = \beta + N(0, s^2). \tag{1}$$

We observe only the pair (b, s). Here, we ignore the difference between the true SD of b given β and the SE estimate s based on the trial data. We define the z-value $z = b/s$ and the signal-to-noise ratio (SNR) $\mathrm{SNR} = \beta/s$. If the outcome is binary, we summarize the effect as a log odds ratio; if it is numerical, we summarize the effect as a standardized mean difference (SMD) which is the difference in the mean outcome between the groups divided by the SD among the participants.

As in References 3 and 4, we select all trials from the CDSR which we could positively identify as RCTs. From each trial we select the comparison that was targeted at efficacy (rather than safety). When there were multiple comparisons (multiple outcomes and/or multiple groups), we select the one that was listed first. In this way, we obtain effect estimates with their SEs from about 20 000 unique RCTs.

We use these data to estimate the joint distribution of the z-value and the SNR. It is a pleasant surprise that this is possible because we cannot observe the SNR directly. The details are as follows. First, we start by estimating the marginal distribution of the z-values. To this end, we use a mixture of four zero-mean normal distributions:

$$z \sim \sum_{i=1}^{4} p_i N\left(0, \sigma_i^2\right).$$

This model is equivalent to assuming that the true SNR follows a mixture of multiple mean zero normal distributions. The variances σ_i^2 and the mixing proportions $p_i, i = 1, 2, 3, 4$ can be estimated using the EM algorithm. We show the histogram of the observed z-values together with the estimated distribution in Figure 1. We verified that this analysis is

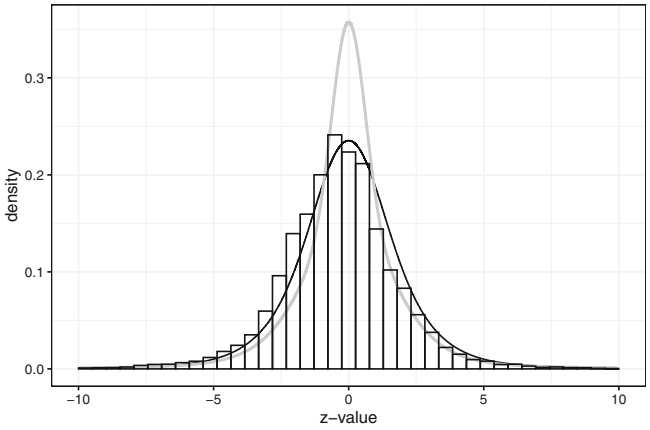

FIGURE 1 The histogram represents the z-values of a little over 20 000 RCTs from the CDSR. The smooth black curve is a mixture of four zero-mean Gaussian components. The smooth gray curve is obtained by subtracting 1 from the variances of these components.

not sensitive to the number of mixing components and the result is very similar when we use 3 or 5 or 6 components in the mixture distribution. We see that the fit is reasonably good, but not perfect. The fit would be better if we did not restrict the components to have zero means. However, this would ultimately result in an asymmetric shrinkage estimator which would not be acceptable in the context of clinical trials.

It follows from (1) that the z-value is the sum of the SNR and independent standard normal noise, which is the result of randomness from observed data in individual trials and not related to SNR, that is,

$$z = \text{SNR} + N(0, 1).$$

As a side note, this observation also implies that the variances in the normal mixture distribution of z must be at least one, that is, $\sigma_i^2 \geq 1$. Hence, we can obtain the marginal distribution of the SNR by removing the standard normal component from the estimated distribution of the z-values. This process is called "deconvolution." It is particularly easy in our case because we are working with mixtures of normal distributions; we can simply subtract 1 from the variances of each of the four mixture components, that is, the distribution of SNR can be estimated by

$$\sum_{i=1}^{4} \hat{p}_i N(0, \hat{\sigma}_i^2 - 1),$$

where $(\hat{p}_i, \hat{\sigma}_i)$ is the estimator for (p_i, σ_i) from the EM algorithm. We show the resulting marginal distribution of the SNR in Figure 1, and report the parameter estimates in Table A1.

Since we are working with a parametric model, it is not difficult to derive the conditional distribution of the SNR given the observed z-value. It is also a mixture of normals and can be estimated by

$$\sum_{i=1}^{4} q_i(z) N(m_i(z), v_i),$$

where $m_i(z) = z(\hat{\sigma}_i^2 - 1)/\hat{\sigma}_i^2$, $v_i = (\hat{\sigma}_i^2 - 1)/\hat{\sigma}_i^2$ are the means and variances of the components. The mixture proportions are

$$q_i(z) = \frac{\hat{p}_i \varphi(z/\hat{\sigma}_i)}{\sum_{j=1}^{4} \hat{p}_j \varphi(z/\hat{\sigma}_i)}, \tag{2}$$

where $\varphi(\cdot)$ is the density function of a standard normal. Now, since $\beta = s \cdot \text{SNR}$, we propose

$$\hat{\beta} = s \cdot \widehat{\mathbb{E}}(\text{SNR}|z) = s \cdot \sum_{i=1}^{4} q_i(z) m_i(z) = b \cdot \sum_{i=1}^{4} q_i(z) \frac{\hat{\sigma}_i^2 - 1}{\hat{\sigma}_i^2}, \tag{3}$$

as an alternative to b for estimating the treatment effect β, see References 3 and 4.

The rationale for the new estimator is that by borrowing information from z-values observed in other studies, we may be able to better estimate the SNR in the current study. It is clear from the definition (3) that $|\hat{\beta}| < |b|$. In other words, $\hat{\beta}$ is a shrinkage estimator.

We can also compute the SD σ of $\hat{\beta}$ and use the interval $\hat{\beta} \pm 1.96\,\sigma$ instead of the usual confidence interval $b \pm 1.96\,s$. Here, σ is s times the SD of the conditional distribution SNR given z. That is,

$$s \cdot \sqrt{\sum_{i=1}^{4} q_i(z)(v_i + m_i(z)^2) - \left(\sum_{i=1}^{4} q_i(z) m_i(z)\right)^2}, \tag{4}$$

yielding a narrower confidence interval for β. All calculations are quite straightforward, and we provide R code in the Appendix.

As we will demonstrate in Section 3, the shrinkage estimator has a substantially better average performance than the conventional counterpart. On the other hand, this improvement may not be realized for each individual trial. The shrinkage in estimating the treatment effect in a study is induced by borrowed information from other studies. When other studies are similar to the study of interest and we have a good idea about the likely SNR value of other studies based on their observed z scores, a bigger improvement can be expected. On the other hand, if (some) other studies are very different from the study of interest in key characteristics, then the advantage of shrinkage may be limited and even vanish completely. Therefore, we may consider to adopt specific shrinkage schemes for a subgroup of more "homogeneous" trials so that a bigger gain of shrinkage can be realized in a larger proportion of trials. For example, one may expect that the distribution of SNR in trials in a disease without any known effective treatment is more concentrated at the mass zero than the average. Consequently, more shrinkage should be applied to such a trial. Operationally, one may consider appropriate grouping of trials of interest based on certain trial characteristics (eg, medical specialty and/or study phase), and apply the proposed adjustment procedure for each subgroup of trials separately. More discussion and analysis can be found in Section 4.

3 | EVALUATION OF THE PERFORMANCE OF THE CONVENTIONAL AND SHRINKAGE ESTIMATORS

3.1 | Setup

In our previous work we have used the primary efficacy results of roughly 20 000 RCTs from the CDSR, while ignoring the fact that the CDSR is actually a collection of systematic reviews. Here we will use this structure to evaluate and compare the performance of the usual estimator b and the shrinkage estimator $\hat{\beta}$ on average across the CDSR.

To this end, we selected all reviews from the CDSR with at least five individual studies. We used the same selection criteria for the individual trials as before, except that we dropped the requirement that a trial must be positively identified as an RCT. Thus we collected the primary efficacy results of 18 226 unique trials from 1625 systematic reviews. We consider the following hierarchical model which is customary for random effects meta-analysis, for example, Reference 12. For the jth individual study in the ith meta-analysis consisting of n_i studies, we assume

$$\beta_{ij} = \mu_i + u_{ij}, \tag{5}$$

$$b_{ij} = \beta_{ij} + \varepsilon_{ij}, \tag{6}$$

where $j = 1, \ldots, n_i$, u_{ij} has the normal distribution with mean zero and variance τ_i^2, and ε_{ij} has the normal distribution with mean zero and variance s_{ij}^2. Moreover, all the u_{ij} and ε_{ij} are assumed to be independent.

We observe the pairs (b_{ij}, s_{ij}) and define the z-values $z_{ij} = b_{ij}/s_{ij}$. We also compute the shrinkage estimators $\hat{\beta}_{ij}$ and their SEs σ_{ij} using the R code provided in the Appendix. In the following, we will present analyses to compare the average performance of the usual and shrinkage estimators.

3.2 | Comparison based on synthetic data

In the first comparison, we generate the synthetic copies of CDSR. Since the true treatment effect for each study is known in generating synthetic CDSR, we may directly evaluate the performance of the two estimators. To construct a synthetic copy of the CDSR, we first conducted standard random effects meta-analyses to obtain estimates $\hat{\mu}_i$ and $\hat{\tau}_i$. We use the function `rma()` from the `R` package `metafor`[12] with default settings. The estimate $\hat{\mu}_i$ is just a weighted average of the individual estimates

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} w_{ij} b_{ij}}{\sum_{j=1}^{n_i} w_{ij}}, \tag{7}$$

where $w_{ij} = 1/(\hat{\tau}_i^2 + s_{ij}^2)$, n_i is the number of individual trials in the ith meta-analysis and $\hat{\tau}_i^2$ is the restricted maximum likelihood estimator of the between study variation. The selection of the estimation method for the meta-analysis is not particularly important for us as the purpose is merely to obtain model parameters which can be used to generate observed data (b, s) whose distribution is similar to its observed counterpart. Next, we perform the following two sampling steps:

1. Sample β_{ij}^* $(j = 1, 2, \ldots, n_i)$ from the normal distribution with mean $\hat{\mu}_i$ and SD $\hat{\tau}_i$.
2. Sample b_{ij}^* from the normal distribution with mean β_{ij}^* and SD s_{ij}. Set $z_{ij}^* = b_{ij}^*/s_{ij}$.

The construction provides us with a simulated set $(\beta_{ij}^*, b_{ij}^*, s_{ij})$ with a similar structure as the original CDSR $(\beta_{ij}, b_{ij}, s_{ij})$. In Figure 2, we compare the distributions of observed b_{ij} and z_{ij} to the those of simulated b_{ij}^* and z_{ij}^*, and note the close agreement between them. We conclude that we not only succeeded in faithfully reproducing the distribution of the estimates b_{ij} but also the relation between the b_{ij} and the SEs s_{ij}.

Finally, we used the `R` code in the Appendix to compute the shrinkage estimators $\hat{\beta}_{ij}^*$ and their SE estimates σ_{ij}^* from the pairs (b_{ij}^*, s_{ij}).

We repeated this data generation and subsequent shrinkage adjustment steps 100 times to reduce the Monte Carlo variation in our final results. This data generation method is essentially a version of the parametric bootstrap based on the random effects models (5) for 1625 meta analyses with all model parameters being their estimators based on original CDSR.

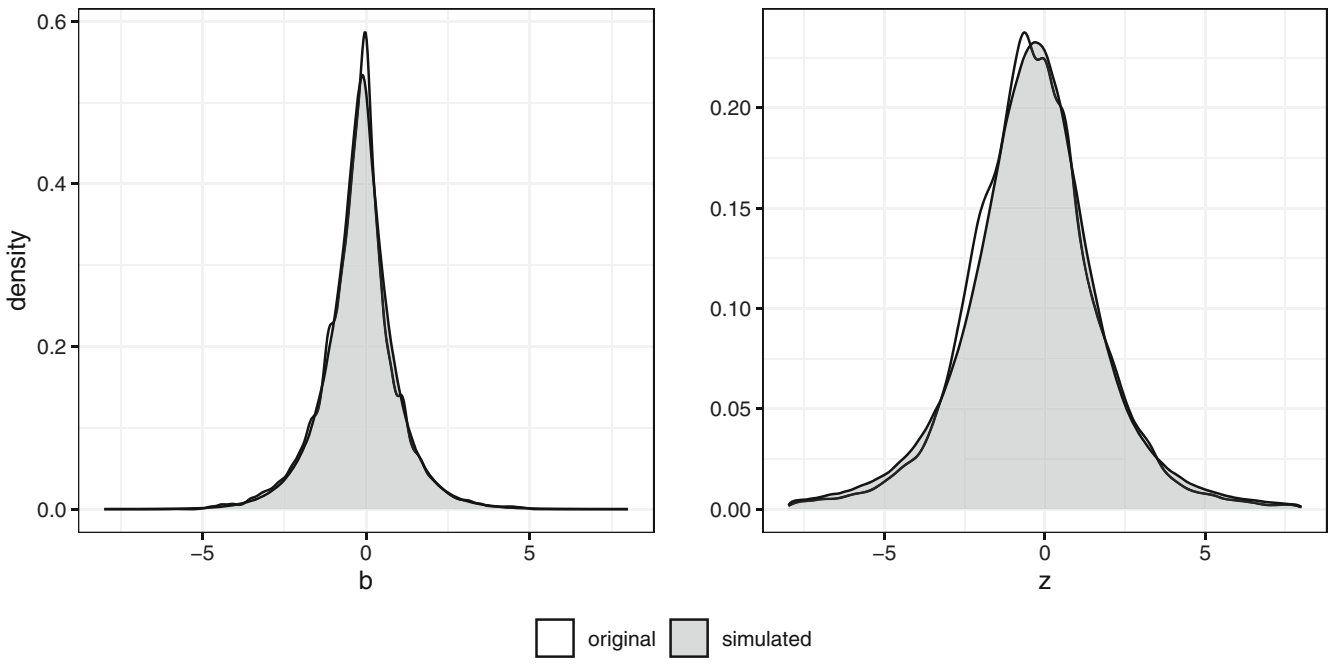

FIGURE 2 Comparison of the distributions of the observed b_{ij} (left) and $z_{ij} = b_{ij}/s_{ij}$ (right) to the simulated b_{ij}^* and $z_{ij}^* = b_{ij}^*/s_{ij}$.

We can now proceed to evaluate and compare the performance of the usual, unbiased estimators b_{ij}^* and the shrinkage estimators $\hat{\beta}_{ij}^*$ on average across the synthetic CDSR, in which the true effect β_{ij}^* is known. With N denoting the total number of individual studies, we define the RMSE, the coverage and the bias of the magnitude as follows.

- RMSE: $\sqrt{\frac{1}{N}\sum_{ij}\left(b_{ij}^*-\beta_{ij}^*\right)^2}$ and $\sqrt{\frac{1}{N}\sum_{ij}\left(\hat{\beta}_{ij}^*-\beta_{ij}^*\right)^2}$.
- Coverage:

$$\frac{1}{N}\sum_{ij}\mathbf{1}\left\{|b_{ij}^*-\beta_{ij}^*|<1.96s_{ij}\right\} \quad \text{and} \quad \frac{1}{N}\sum_{ij}\mathbf{1}\left\{|\hat{\beta}_{ij}^*-\beta_{ij}^*|<1.96\sigma_{ij}^*\right\}.$$

- Bias of the magnitude: $\frac{1}{N}\sum_{ij}\left(|b_{ij}^*|-|\beta_{ij}^*|\right)$ and $\frac{1}{N}\sum_{ij}\left(|\hat{\beta}_{ij}^*|-|\beta_{ij}^*|\right)$.

We report these performance measures in Table 1 and note the superior performance of the shrinkage estimator. The RMSE is reduced from 0.73 to 0.54, which implies a relative gain in efficiency of $(0.73^2-0.54^2)/0.73^2=0.45$. So, in a sense, this improvement corresponds to almost doubling the sample size. Moreover, the shrinkage does not lead to severe downward bias. In fact, on average over the CDSR the unbiased estimator tends to severely overestimate the magnitude of the effect. This is to be expected from Jensen's inequality, which in our case states that $\mathbb{E}(|b_{ij}|)>|\mathbb{E}(b_{ij})|=|\beta_{ij}|$. The bias is large when the absolute value of the SNR is small. This is often the case among the trials in the CDSR and therefore, on average over the CDSR, the bias of the magnitude is substantial. The shrinkage effectively corrects this. Finally, the slight undercoverage of the interval around the shrinkage estimator is due to the fact that we use a normal approximation to the conditional distribution of the SNR given the z-value. In Section 4, we break this table down by medical specialty.

We decided not to report the average bias of the two estimators in Table 1. The average bias of the unbiased estimator is of course zero. Now, as can be seen from Figure 1, the distribution of the observed z-values is nearly symmetric around zero. Since the shrinkage estimator shrinks towards zero, its bias on average over the CDSR will necessarily also be close to zero. Thus, the average bias is not useful for comparing the performance of the two estimators.

In Table 2, we report the performance measures conditional on statistical significance at the 5% level by restricting the averages to the cases where $|z_{ij}^*|>1.96$. We note that the bias in the magnitude of the usual estimator has become even larger due to the winner's curse. Also note the substantial undercoverage of the usual confidence interval. As can be seen from its definition (3), the shrinkage estimator is computed conditionally on the observed z-value. Therefore it is not affected by the winner's curse.

3.3 | Comparison based on cross-validation

In generating the synthetic CDSR, the true treatment effect for each study is known allowing direct evaluation of the performance of relevant estimators. The evaluation, however, depends on the model and parameters used to generate the

TABLE 1 Performance of the estimators.

Estimator	RMSE	Bias of the magnitude	Coverage	Ave. width of the CI
Unbiased	0.73	0.21	0.95	2.34
Shrinkage	0.54	−0.05	0.94	1.92

TABLE 2 Performance of the estimators conditional on statistical significance, that is, $|b_{ij}^*/s_{ij}|>1.96$.

Estimator	RMSE	Bias of the magnitude	Coverage	Ave. width of the CI
Unbiased	0.79	0.37	0.90	1.65
Shrinkage	0.55	0.05	0.95	1.62

true treatment effect and synthetic CDSR. The close agreement of the original and simulated data in Figure 2 provides partial support that the observed gain of the shrinkage estimators in synthetic the CDSR is likely to be real in practice. In this section, we provide additional support without simulating synthetic the CDSR.

We naturally think of both b_{ij} and $\hat{\beta}_{ij}$ as estimators of β_{ij}. However, since the β_{ij} are normally distributed with mean μ_i, we can *also* view both b_{ij} and $\hat{\beta}_{ij}$ as estimators of μ_i. As such, we can compare their performance by their mean squared estimation errors,

$$\mathrm{MSE}_b = \frac{1}{N}\sum_{ij}\left(b_{ij}-\mu_i\right)^2 \quad \text{and} \quad \mathrm{MSE}_{\hat{\beta}} = \frac{1}{N}\sum_{ij}\left(\hat{\beta}_{ij}-\mu_i\right)^2. \tag{8}$$

Since b_{ij} is an unbiased estimator of μ_i, it follows that

$$\mathbb{E}\left[(b_{ij}-\mu_i)^2|\mu_i\right] = s_{ij}^2 + \tau_i^2. \tag{9}$$

So, we can estimate the mean squared error of the b_{ij} directly as

$$\frac{1}{N}\sum_{ij}\left(s_{ij}^2+\hat{\tau}_i^2\right) = 0.74. \tag{10}$$

Since the $\hat{\beta}_{ij}$ are not unbiased estimators of the μ_i, we cannot estimate the mean squared error of the $\hat{\beta}_{ij}$ in a similar way. We will therefore use a different approach by focusing on the difference in mean squared errors,

$$\mathrm{MSE}_b - \mathrm{MSE}_{\hat{\beta}} = \frac{1}{N}\sum_{ij}\left[(b_{ij}-\mu_i)^2 - (\hat{\beta}_{ij}-\mu_i)^2\right]. \tag{11}$$

The μ_i are not observed, but, remarkably, there is a direct way to estimate this difference. We start by constructing a third estimator of μ_i which is unbiased and independent of both b_{ij} and $\hat{\beta}_{ij}$. We leave out study j from meta-analysis i and re-run the random effects meta-analysis on the remaining $n_i - 1$ studies to obtain an estimate $\hat{\mu}_i^{(-j)}$ of the average effect μ_i. The estimate $\hat{\mu}_i^{(-j)}$ is a weighted average of the $n_i - 1$ estimates from the individual studies without the jth one,

$$\hat{\mu}_i^{(-j)} = \frac{\sum_{k\neq j} w_{ik} b_{ik}}{\sum_{k\neq j} w_{ik}}, \tag{12}$$

where $w_{ik} = 1/\left(\hat{\tau}_{i(-j)}^2 + s_{ik}^2\right)$ and $\hat{\tau}_{i(-j)}^2$ is an estimator of τ_i^2 based on $n_i - 1$ studies excluding the jth study. This estimator is unbiased for μ_i because the individual study estimates are. Moreover, it is independent of both b_{ij} and $\hat{\beta}_{ij}$ because it is based on a different set of studies. (To be very precise, the amount of shrinkage is derived from all the RCTs in the CDSR, so in that sense $\hat{\beta}_{ij}$ does depend a little bit on the other studies in the ith meta-analysis. But this dependence is very slight and can be safely ignored.)

The unbiasedness of the $\hat{\mu}_i^{(-j)}$ and their independence of b_{ij} and $\hat{\beta}_{ij}$ imply the following equality

$$\mathbb{E}\left[(b_{ij}-\mu_i)^2 - (\hat{\beta}_{ij}-\mu_i)^2\right] = \mathbb{E}\left[\left(b_{ij}-\hat{\mu}_i^{(-j)}\right)^2 - \left(\hat{\beta}_{ij}-\hat{\mu}_i^{(-j)}\right)^2\right], \tag{13}$$

which implies that

$$\frac{1}{N}\sum_{ij}\left[(b_{ij}-\mu_i)^2 - (\hat{\beta}_{ij}-\mu_i)^2\right] \approx \frac{1}{N}\sum_{ij}\left[\left(b_{ij}-\hat{\mu}_i^{(-j)}\right)^2 - \left(\hat{\beta}_{ij}-\hat{\mu}_i^{(-j)}\right)^2\right]$$

for large N. We provide a proof of this equality in the Appendix. Therefore, the right-hand side is directly observable from the original CDSR and can be viewed as a leave-one-out cross-validation estimate of the left-hand side of the equality, which coincides with (11). Based on our data, this leave-one-out cross validation estimate turns out to be 0.33. Recalling that we estimated $\mathrm{MSE}_b = 0.74$, we can now estimate $\mathrm{MSE}_{\hat{\beta}} = 0.74 - 0.33 = 0.41$. This shows that on average, the

shrinkage estimator provides a relative efficiency gain over the unbiased estimator of 55% for estimating the μ_i. Similar to the improvement in estimating the true effect of individual studies, this improvement also corresponds to more than doubling the sample size.

Remark. Circling back to our random effects model in Section 3.2, we can simulate multiple synthetic copies of CDSR from the model, and compute the corresponding mean squared errors

$$\frac{1}{N}\sum_{ij}\left(b^*_{ij}-\hat{\mu}_i\right)^2 \quad \text{and} \quad \frac{1}{N}\sum_{ij}\left(\hat{\beta}^*_{ij}-\hat{\mu}_i\right)^2. \tag{14}$$

These turns out to be 0.74 and 0.42, respectively, which is in close agreement to the results obtained above from the leave-one-out analysis. This strengthens our confidence in both the synthetic CDSR model and the leave-one-out cross validation analysis.

3.4 | A visual explanation

In this article, we have used the terms "bias," "RMSE," and "coverage" *not* in the usual (frequentist) sense where we fix a particular value for the true effect β. Instead, we average over the effects that occur in the CDSR. Tables 1 and 2 clearly show that the shrinkage estimator performs much better than the usual unbiased estimator on average across the CDSR. The reason for this superiority is that the shrinkage estimator uses shared information between the trials. This is the well-known "Stein effect."[13-15]

We illustrate the Stein effect in Figure 3 where we note superior performance of the shrinkage estimator (top panel) at the most common values of the true effect (bottom panel). We also see that the estimator shrinks too much when the true effect is very large. Fortunately, very large effects are rare as can be seen in the bottom panel.

Figure 3 may appear to suggest that the shrinkage estimator is likely to overshrink when the estimated effect is large, but that is no so. In fact, the shrinkage is particularly effective in that case. In Figure 4, we plot the (synthetic) estimated effects b^*_{ij} versus the difference of the squared errors of the two estimators, that is, $\left(b^*_{ij}-\beta^*_{ij}\right)^2-\left(\hat{\beta}^*_{ij}-\beta^*_{ij}\right)^2$. The "loess" regression curve represents the conditional expectation of this difference given the estimated effect.[16] We see that this is always in favor of the shrinkage estimator, but especially when the observed effect is large.

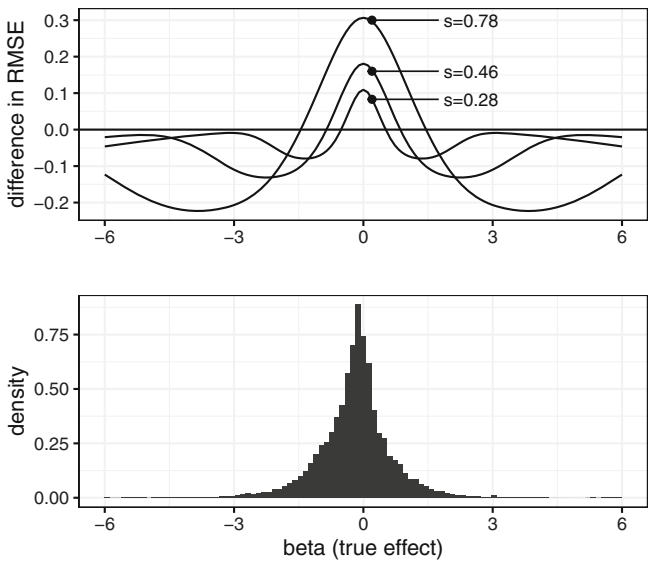

FIGURE 3 Top panel: The difference in RMSE between the unbiased estimator and the shrinkage estimator as a function of the true effect β when the SE s of the unbiased estimator is set to its quartiles across the CDSR (0.28, 0.46, and 0.78). A positive difference indicates superior performance of the shrinkage estimator. Bottom panel: The distribution of the true effect β in a generated synthetic copy of CDSR.

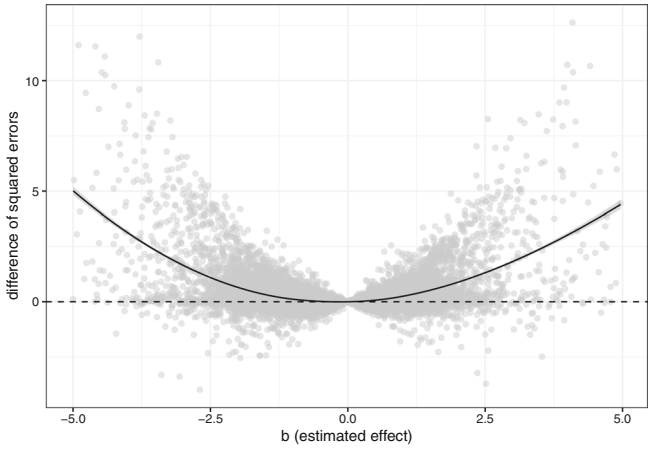

FIGURE 4 The estimated effects vs the difference of the squared errors of the unbiased estimator and the shrinkage estimator from a synthetic copy of CDSR. The smooth curve is the loess regression. Positive values favor the shrinkage estimator.

The "price to pay" for the Stein effect is that the shrinkage estimator $\hat{\beta}$ is biased in the frequentist sense, that is, for any fixed value of the true parameter β. Specifically, the shrinkage towards zero causes the magnitude of the estimator to be systematically too small. However, we have already seen in Tables 1 and 2 that when we average over the β_{ij} in the CDSR, the bias of the magnitude of the shrinkage estimator is actually much *less* than that of the usual estimator. So, we obtain a reduction of both the variance and the bias of the magnitude of the effect on average across the CDSR.

Of course, superior performance on average over the CDSR does not guarantee superior performance in any particular trial. We can see in Figure 4 that the shrinkage estimator is sometimes further from the truth than the unbiased estimator. From Figure 3 we know that the shrinkage estimator will perform poorly when the true effect β is very large. Very large effects are rare, but even more importantly, we do not know if the true effect is large. This is the crux: clearly, we should use the unbiased estimator if we can somehow *recognize* that it will perform best. Otherwise, it seems sensible to prefer the shrinkage estimator because it may be *expected* to perform better, as we can see in Figure 4.

4 | STRATIFICATION BY MEDICAL FIELD

It is not obvious that the superior performance of the shrinkage estimator on average across the CDSR is relevant for the inference about a particular trial. For any particular trial there will be additional information that sets it apart from all the other trials, such as the trial design, the medical field, the outcome, whether it was run by a pharmaceutical company or a university hospital and so on and so forth.

The fact that we have additional information about a particular trial does not mean that the shrinkage estimator is not useful. On the contrary, such information can be used to improve the shrinkage estimator when appropriately used. For example, the trials in the CDSR are classified into 19 medical specialties. We can construct shrinkage estimators within each of these specialties. Specifically, for each medical specialty separately, we can re-estimate the joint distribution of the z-value and SNR and computed shrinkage estimators and their SEs. We refer these new shrinkage estimators as "local" shrinkage estimators. The resulting shrinkage would be different for trials in different medical specialties due to different estimates for the conditional distribution of SNR given z. We can then evaluate and compare the performance of the usual estimator, the shrinkage estimator based on entire CDSR, and "local" shrinkage estimators within these specialties. We show the results of this stratification in Figure 5 where we sorted the specialties by the number of trials, see also Table C1 in the Appendix. We find that there is very little difference between the performance of local and global shrinkage.

With information on additional trial characteristics, we may construct even more "local" shrinkage estimators within smaller subgroup of trials sharing the same characteristics, such as trials from the same pharmaceutical company with an excellent track record in conducting successful trials. The estimated shrinkage is expected to be more tuned to individual trials in that subgroup and might bring bigger improvement in estimation accuracy. In particular, the shrinkage may be more flexible and is not always towards zero, since the components of the mixture distribution for SNR may not be mean zero. However, there is a practical limitation of this approach, since sufficient number of trials is needed to construct a

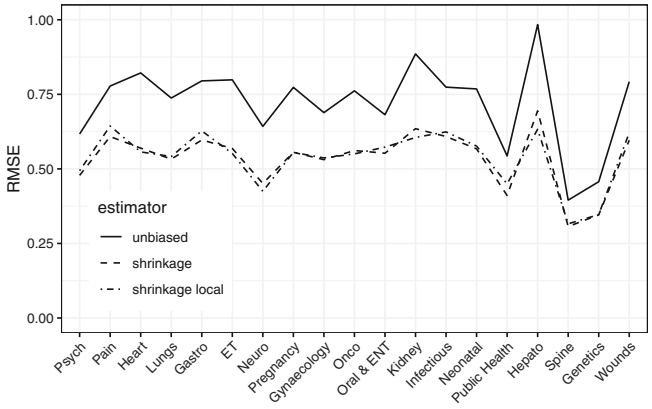

FIGURE 5 Root mean squared error of the three different estimators, compare Table C1.

good enough estimate for the joint distribution of z and SNR. Otherwise, the gain of the shrinkage may be offset by the inaccuracy in estimating the conditional expectation $\mathbb{E}(\text{SNR}|z)$.

Besides the practical issue of the availability of sufficient relevant information, there is also the conceptual issue of deciding which information should be used, and, ultimately, which studies are relevant for estimating the distribution of SNR. This is essentially a subjective choice, which could jeopardize the validity of the inference by opening the door to the "garden of forking paths."[17]

5 | DISCUSSION

We can represent "the essence" of a clinical trial by a set of three numbers (β, b, s), where β is the primary efficacy parameter and b is an unbiased, normally distributed estimator with SE s. In Reference 3 we estimated the joint distribution of the z-value $z = b/s$ and the SNR $\text{SNR} = \beta/s$ across approximately 20 000 trials from the Cochrane database (CDSR). In References 3 and 4, we proposed a shrinkage estimator $\hat{\beta} = s \cdot \widehat{\mathbb{E}}(\text{SNR}|z)$ as an alternative to the unbiased estimator b. We expect that the new shrinkage estimator is more accurate than its unbiased counterpart from the study alone and might be used for future study design to avoid insufficient sample size due to overoptimistic hypothesis on the treatment effect.

It is likely that the empirical Bayes estimator $\widehat{\mathbb{E}}(\beta|b, s)$ would be a better estimator than $\hat{\beta}$ because there is more information in the pair (b, s) than in their ratio z. However, to compute $\widehat{\mathbb{E}}(\beta|b, s)$, we would need the full joint distribution of (β, b, s), which is much more difficult to estimate than the joint distribution of z-value and the SNR.

The goal of this article is to evaluate the performance of $\hat{\beta}$, and compare it to the usual, unbiased estimator. We find that, on average across the CDSR, the shrinkage estimator is much superior to the unbiased estimator in terms of mean squared error, exaggeration and coverage. The improvement is considerable, and continues to hold when we stratify by medical specialty.

The question remains: Is the performance "on average across the CDSR" relevant? Clearly, the CDSR is not a random sample from the population of all clinical trials. However, all trials have in common that they try to obtain a sufficiently precise estimate of the treatment effect within the constraints of time, money, and the availability of subjects. By comparing the shrinkage estimator to the unbiased estimator across the CDSR, we see the performance gain under these shared circumstances. An even more difficult follow-up question is: Is the performance on average across the CDSR relevant for the inference about a particular trial? We believe it is! It is well-known that trials tend to have low SNR (they are often "underpowered") and it would be irresponsible to ignore that.

While the unbiased estimator and its SE (or confidence interval) should always be reported, we would argue that the shrinkage estimator is also important to aid the interpretation and to guard against exaggeration. So, we suggest that the primary result of a trial is reported as follows (the numbers are taken from an example in Reference 3):

> The hazard ratio was estimated at 0.75 with 95% confidence interval of (0.55, 1.02). However, it has been established that many trials have a low signal-to-noise ratio, which can lead to upward bias in the estimate of the hazard ratio. If we apply this general information to our particular trial, the hazard ratio estimate becomes 0.84 with interval (0.62, 1.07).

Finally, after demonstrating the superior performance of the proposed shrinkage estimation procedure, we stress that it is neither a complete replacement of the trial-specific analysis, nor of a well-conducted meta-analysis of a collection of high-quality studies examining the same treatment effect. Furthermore, if feasible, it is also desirable to examine the treatment effect estimates from a group of "similar" studies. Even simple summary such as the range of observed treatment effects from similar studies may provide additional insight to the true treatment effect in the current study.

DATA AVAILABILITY STATEMENT

We provide an online supplement with R code that reproduces all the results in this article from publicly available data.

ORCID

Erik W. van Zwet https://orcid.org/0000-0001-5537-3179
Lu Tian https://orcid.org/0000-0002-5893-0169

REFERENCES

1. Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. *BMJ Open*. 2021;11(9):e045942.
2. Davey J, Turner RM, Clarke MJ, Higgins J. Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):1-11.
3. van Zwet E, Schwab S, Senn S. The statistical properties of RCTs and a proposal for shrinkage. *Stat Med*. 2021;40(27):6107-6117.
4. van Zwet E, Schwab S, Greenland S. Addressing exaggeration of effects from single RCTs. *Significance*. 2021;18(6):16-21.
5. van Zwet E, Goodman S. How large should the next study be? Predictive power and sample size requirements for replication studies. *Stat Med*. 2022;41:3090-3101.
6. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365.
7. Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. *R Soc Open Sci*. 2017;4(2):160254.
8. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648.
9. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9(6):641-651.
10. van Zwet E, Cator E. The significance filter, the winner's curse and the need to shrink. *Stat Neerl*. 2021;75:437-452.
11. Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. *J Am Stat Assoc*. 1975;70(350):311-319.
12. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
13. Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol 1. Berkeley, CA: University of California Press; 1956:197-206.
14. Efron B, Morris C. Stein's estimation rule and its competitors—an empirical Bayes approach. *J Am Stat Assoc*. 1973;68(341):117-130.
15. Stigler SM. The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Stat Sci*. 1990;5:147-155.
16. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*. 1979;74(368):829-836.
17. Gelman A, Loken E. *The Garden of Forking Paths: why Multiple Comparisons Can be a Problem, Even when there Is no "Fishing Expedition" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Time*. Vol 348. New York: Department of Statistics, Columbia University; 2013:1-17.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: van Zwet EW, Tian L, Tibshirani R. Evaluating a shrinkage estimator for the treatment effect in clinical trials. *Statistics in Medicine*. 2024;43(5):855-868. doi: 10.1002/sim.9992

APPENDIX A. COMPUTING THE SHRINKAGE ESTIMATOR

In Reference 3, we estimated the distribution of the SNR β/s as a normal mixture of four zero-mean components which we specify in Table A1.

We compute the shrinkage estimator $\hat{\beta}$ and its SD σ from the unbiased estimator b and its SD s with the following R code:

TABLE A1 Estimated 4-part zero-mean normal mixture distributions of the SNR, from Reference 3.

	comp.1	comp.2	comp.3	comp.4
Proportions	0.32	0.31	0.30	0.07
Means	0	0	0	0
Std. devs.	0.61	1.42	2.16	5.64

```
shrinkage <- function(b,s) {
  z <- b/s
  p <- c(0.32,0.31,0.3,0.07)      # from Table 3
  sigma <- c(0.61,1.42,2.16,5.64)
  sigma2 <- sigma^2
  q <- p*dnorm(z,0,sqrt(sigma2+1))
  q <- q/sum(q)                   # conditional mixing probs
  pm <- b*sigma2/(sigma2+1)       # conditional means
  pv <- s^2*sigma2/(sigma2+1)     # conditional variances
  data.frame(q,pm,pv)
}
```

For example, if we observe $b = 0.4$ and $s = 0.3$ then the usual 95% confidence interval is −0.2 to 1.0. We can compute the shrinkage estimator as follows:

```
shrink <- shrinkage(b=0.4,s=0.3)
betahat <- sum(shrink$q * shrink$pm)
```

We find $\hat{\beta} = 0.23$. We can compute SD σ of the mixture distribution as follows

```
pm2 <- sum(shrink$q * shrink$pm^2)
ps2 <- sum(shrink$q * shrink$pv)
sigma <- sqrt(ps2 + pm2 - betahat^2)
```

We find $\sigma = 0.25$, so the 95% interval is −0.25 to 0.72.

APPENDIX B. EQUATION (13)

We can motivate Equation (13) with the following proposition.

Proposition 1. *Consider three estimators T_0, T_1, and T_2 of a parameter θ. Suppose that, conditionally on θ, T_0 is unbiased and independent of T_1 and T_2. Then*

$$\mathbb{E}(T_1 - T_0)^2 - \mathbb{E}(T_2 - T_0)^2 = \mathbb{E}(T_1 - \theta)^2 - \mathbb{E}(T_2 - \theta)^2, \tag{B1}$$

where the expectations are with respect to arbitrary distributions of T_0, T_1, T_2, and θ (as long as the expectations are well-defined and finite).

Proof.

$$\begin{aligned}\mathbb{E}(T_1 - T_0)^2 &= \mathbb{E}(\mathbb{E}((T_1 - T_0)^2|\theta)) \\ &= \mathbb{E}[\mathbb{E}((T_1 - \theta)^2|\theta) - 2\mathbb{E}((T_1 - \theta)(T_0 - \theta)|\theta) + \mathbb{E}((T_0 - \theta)^2|\theta)].\end{aligned}$$

Conditionally on θ, T_0 is unbiased and independent of T_1.Therefore the cross term is zero and hence

$$\mathbb{E}(T_1 - T_0)^2 = \mathbb{E}(T_1 - \theta)^2 + \mathbb{E}(T_0 - \theta)^2.$$

The same argument holds with T_2 instead of T_1, and the claim follows. ■

Equation (13) follows from the proposition by taking $T_0 = \hat{\mu}_i^{-j}$, $T_1 = b_{ij}$, $T_2 = \hat{\beta}_{ij}$, and $\theta = \mu_i$.

APPENDIX C. STRATIFICATION BY MEDICAL FIELD

TABLE C1 Performance of the estimators.

Specialty	*n*	Estimator	RMSE	Bias of the magnitude	Coverage	Width of the full CI
Psych	2609	Unbiased	0.62	0.16	0.95	1.98
Psych	2609	Shrinkage	0.48	−0.08	0.94	1.64
Psych	2609	Shrinkage local	0.49	−0.03	0.95	1.71
Pain	1757	Unbiased	0.78	0.17	0.95	2.55
Pain	1757	Shrinkage	0.61	−0.13	0.93	2.14
Pain	1757	Shrinkage local	0.64	−0.03	0.95	2.27
Heart	1558	Unbiased	0.82	0.29	0.95	2.59
Heart	1558	Shrinkage	0.57	0.02	0.95	2.10
Heart	1558	Shrinkage local	0.56	0.00	0.95	2.06
Lungs	1274	Unbiased	0.74	0.21	0.95	2.30
Lungs	1274	Shrinkage	0.53	−0.05	0.95	1.88
Lungs	1274	Shrinkage local	0.54	−0.03	0.95	1.91
Gastro	1194	Unbiased	0.80	0.21	0.95	2.60
Gastro	1194	Shrinkage	0.60	−0.08	0.94	2.15
Gastro	1194	Shrinkage local	0.63	−0.01	0.95	2.26
ET	1176	Unbiased	0.80	0.27	0.95	2.61
ET	1176	Shrinkage	0.57	−0.01	0.95	2.12
ET	1176	Shrinkage local	0.55	−0.05	0.95	2.04
Neuro	1152	Unbiased	0.64	0.21	0.95	2.01
Neuro	1152	Shrinkage	0.45	−0.01	0.95	1.63
Neuro	1152	Shrinkage local	0.42	−0.06	0.94	1.52
Pregnancy	1146	Unbiased	0.77	0.24	0.95	2.49
Pregnancy	1146	Shrinkage	0.56	−0.03	0.95	2.03
Pregnancy	1146	Shrinkage local	0.56	−0.03	0.95	2.03
Gynecology	870	Unbiased	0.69	0.19	0.95	2.27
Gynecology	870	Shrinkage	0.53	−0.07	0.94	1.88
Gynecology	870	Shrinkage local	0.54	−0.06	0.94	1.90
Onco	706	Unbiased	0.76	0.23	0.95	2.45
Onco	706	Shrinkage	0.56	−0.03	0.95	2.00
Onco	706	Shrinkage local	0.55	−0.06	0.95	1.95

(Continues)

TABLE C1 (Continued)

Specialty	*n*	Estimator	RMSE	Bias of the magnitude	Coverage	Width of the full CI
Oral and ENT	647	Unbiased	0.68	0.13	0.95	2.12
Oral and ENT	647	Shrinkage	0.55	−0.13	0.93	1.79
Oral and ENT	647	Shrinkage local	0.57	−0.04	0.95	1.90
Kidney	630	Unbiased	0.89	0.31	0.95	2.98
Kidney	630	Shrinkage	0.63	−0.01	0.96	2.41
Kidney	630	Shrinkage local	0.61	−0.07	0.94	2.27
Infectious	621	Unbiased	0.77	0.20	0.95	2.45
Infectious	621	Shrinkage	0.61	−0.08	0.94	2.05
Infectious	621	Shrinkage local	0.62	−0.05	0.95	2.11
Neonatal	550	Unbiased	0.77	0.23	0.95	2.55
Neonatal	550	Shrinkage	0.57	−0.05	0.95	2.09
Neonatal	550	Shrinkage local	0.58	−0.03	0.95	2.12
Public Health	524	Unbiased	0.54	0.13	0.95	1.55
Public Health	524	Shrinkage	0.41	−0.05	0.93	1.30
Public Health	524	Shrinkage local	0.45	0.02	0.95	1.40
Hepato	517	Unbiased	0.98	0.41	0.95	3.45
Hepato	517	Shrinkage	0.69	0.05	0.96	2.78
Hepato	517	Shrinkage local	0.63	−0.11	0.93	2.37
Spine	512	Unbiased	0.40	0.09	0.95	1.27
Spine	512	Shrinkage	0.31	−0.06	0.94	1.06
Spine	512	Shrinkage local	0.32	−0.04	0.94	1.10
Genetics	423	Unbiased	0.46	0.11	0.95	1.38
Genetics	423	Shrinkage	0.35	−0.06	0.93	1.16
Genetics	423	Shrinkage local	0.35	−0.05	0.93	1.15
Wounds	360	Unbiased	0.79	0.19	0.95	2.55
Wounds	360	Shrinkage	0.60	−0.10	0.94	2.10
Wounds	360	Shrinkage local	0.62	−0.04	0.95	2.20

Abbreviations: ET, emergency and trauma; Gastro, gastroenterology; Genetics, genetics and endocrinology; Gynecology, gynecology and urology; Heart, heart and hypertension; Hepato, hepato-biliary; Infectious, infectious diseases; kidney, kidney and transplant; Oral and ENT, oral health, Eyes and ENT; Pain, anesthesia and pain; Pregnancy, pregnancy and childbirth; Psych, Psychiatry and Mental Health; Pub Health, public health and work; Spine, spine and muscles; Wounds, skin and wounds.